RESEARCH PAPER

Digital Remote Patient Monitoring: An Artificial Intelligence-Based Clinical Decision Support System for Hypertension Management

Saniya Raheen Patel[1], Vandana C. Bagal[2], Abhijit Banubakode[3], Bharat Naiknaware[4], Nazneen Akhter[5]

[1,4]Dr. G.Y. Pathrikar College of C.S. & I.T., MGM University, Chh. Sambhajinagar, India.
[2] K. K.Wagh Institute of Engineering Education and Research, Nashik,Maharashtra, India.,
[3]Pimplri Chinchwad University, Pune, Maharashtra, India
[5]*IRC Technical Institute, Chh. Sambhajinagar, India.*
[1]saniapatel72.sp@yahoo.com , [2]vcbagal@kkwagh.edu.in , [3]abhijitsiu@gmail.com , [4]bbharat.naiknaware@gmail.com , [5]getnazneen@gmail.com

corresponding author: getnazneen@gmail.com

Abstract

Hypertension remains a leading modifiable risk factor for cardiovascular disease worldwide, yet conventional episodic clinic-based blood pressure (BP) monitoring frequently fails to capture the temporal dynamics of disease progression. This paper presents and empirically evaluates an artificial intelligence (AI)-based Clinical Decision Support System (CDSS) built on a synthetically generated, clinically grounded remote patient monitoring (RPM) dataset comprising 573 simulated patients and 297,809 longitudinal BP/heart-rate readings collected over a six-month period. We construct a patient-level feature representation that fuses static demographic/clinical risk factors with longitudinal vital-sign summary statistics, and evaluate Logistic Regression and Random Forest classifiers across six experiments: (1) demographic-only risk stratification, (2) fused demographic-plus-vitals risk stratification, (3) early-window (30-day) prediction of disease progression propensity, (4) robustness under non-random (MNAR) missing data, (5) subgroup performance analysis across age, sex, and BMI strata, and (6) reading-level hypertensive crisis (Stage-2) alert detection compared against a standard clinical threshold rule. The fused model achieved a macro F1-score of 0.715 (accuracy 77.8%) for three-class clinical risk classification, and the Random-Forest-based alert classifier achieved perfect discrimination (F1 = 1.00, AUC ≈ 1.00) for Stage-2 hypertensive readings versus 0.644 F1 for a fixed-threshold rule, reducing the false-alarm rate from 9.9% to 0%. We further show that adherence-aware feature engineering partially recovers performance lost to non-random missingness, and that model accuracy varies meaningfully across demographic subgroups — underscoring the need for subgroup-aware validation before clinical deployment. These findings demonstrate the technical feasibility and clinical relevance of AI-driven RPM-based hypertension CDSS while highlighting concrete directions for fairness auditing and real-world validation.

Keywords: remote patient monitoring; hypertension; clinical decision support system; machine learning; digital health; missing data; risk stratification

How to cite this article: Patel SR, Bagal VC, Banubakode A, Naiknaware B, Akhter N. Digital remote patient monitoring: an artificial intelligence-based clinical decision support system for hypertension management. Int J Drug Deliv Technol. 2026;16(75s): 645-651. DOI: 10.25258/ijddt.16.75s.76

1. Introduction

Hypertension affects an estimated 1.28 billion adults globally [1] and is the single largest contributor to cardiovascular mortality [2], yet nearly half of affected individuals are unaware of their condition and even fewer achieve adequate control through periodic clinical visits alone. The rise of consumer-grade and clinical-grade digital blood pressure devices has enabled Remote Patient Monitoring (RPM) [3] , in which patients transmit BP, heart rate, and related vitals from home on a near-continuous basis. RPM generates orders of magnitude more data than episodic clinic visits, creating both an opportunity and a challenge [4]: clinicians cannot manually review thousands of daily readings per patient panel, motivating the need for Artificial Intelligence (AI)-based Clinical Decision Support Systems (CDSS) capable of triaging, stratifying, and forecasting patient risk automatically [5].

This paper investigates the design and empirical performance of such a system using a synthetically generated but clinically grounded RPM dataset. Synthetic data allows controlled evaluation of modeling strategies — including realistic complications such as non-random missing data driven by patient adherence — without the privacy constraints of real patient records, while remaining structured to reflect plausible clinical relationships between demographics, comorbidities, and BP trajectories. Specifically, we address the following research questions:

- RQ1: How accurately can a CDSS stratify patient hypertension risk using static demographic/clinical features versus longitudinal RPM vitals, and does fusing both sources improve performance?

- RQ2: Can early-window RPM data (the first 30 days of monitoring) predict a patient's longer-term disease progression trajectory?
- RQ3: How robust is model performance to non-random missing data patterns driven by patient monitoring adherence, and can adherence-aware features mitigate performance loss?
- RQ4: Does model performance vary meaningfully across demographic subgroups (age, sex, BMI)?
- RQ5: Can a machine-learned alert classifier outperform a fixed clinical threshold rule for detecting Stage-2 (hypertensive crisis-level) readings, in terms of precision and false-alarm burden?

We answer these questions using a reproducible experimental pipeline over four linked data sources: a longitudinal vitals stream, a matched non-random-missingness (MNAR) variant of that stream, a static demographic/clinical risk-factor profile, and a disease-progression metadata file. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset and feature engineering; Section 4 describes the experimental setup; Section 5 presents results; Section 6 discusses clinical and technical implications; Section 7 outlines limitations; and Section 8 concludes with directions for future work.

2. Related Work

2.1 Remote Patient Monitoring for Hypertension

Clinical guidelines increasingly recognize out-of-office BP measurement [6] including home BP monitoring and ambulatory BP monitoring as more predictive of cardiovascular outcomes than office-based readings alone [7], motivating the shift toward continuous RPM-based management. RPM programs have been associated with improved BP control rates in primary-care settings [8], but existing implementations frequently rely on simple rule-based alerting (e.g., static threshold crossings) [9] rather than adaptive, learned risk models.

2.2 AI-Based Clinical Decision Support

Machine learning approaches to hypertension risk stratification have explored a range of model families, from interpretable linear and tree-based models to deep sequence models for time-series vitals [10]. A recurring theme in this literature is the trade-off between model complexity/accuracy and clinical interpretability [11], since CDSS outputs must be explainable to treating clinicians to support trust and adoption. Ensemble tree-based methods such as Random Forests are commonly favored in clinical ML applications for their strong tabular performance and natural feature-importance interpretability.

2.3 Missing Data in Digital Health

Data missingness in RPM streams is rarely missing completely at random (MCAR) [12]; it is frequently missing not at random (MNAR), correlated with patient engagement, symptom burden, or device adherence. This has direct implications for CDSS reliability [13], since naive imputation strategies that ignore the missingness mechanism can bias downstream risk estimates. Our experimental design explicitly incorporates an MNAR-structured variant of the dataset to evaluate this concern [14].

3. Dataset and Methods

3.1 Data Sources

The study uses four linked, patient-level-keyed synthetic data files covering 573 simulated patients:

- Longitudinal vitals stream (297,809 readings, Jan–Jun 2025): timestamped systolic BP (SBP), diastolic BP (DBP), heart rate (HR), mean arterial pressure (MAP), pulse pressure, categorical BP stage (normal / elevated / stage_1 / stage_2), and a derived risk_state (low / medium / high).
- MNAR vitals variant: an identical structure with 7.4% of readings set to missing via a non-random mechanism; missingness is markedly higher for readings that would otherwise be labeled low risk (10.0%) than high risk (1.9%), consistent with reduced monitoring engagement when patients feel well.
- Static demographic/clinical profile: age, sex, menopause status, height, weight, BMI (and category), smoking status, diet quality, physical activity level, diabetes flag, prior cardiac event flag, a four-level hypertension_status label, and a three-level clinical_risk label (Low / Moderate / High).
- Progression profile: a continuous progression_propensity score (0–1), a discretized propensity_band (low / medium / high), and a baseline_adherence level (high / medium) intended to govern monitoring engagement over time.

3.2 Feature Engineering

Longitudinal readings were aggregated to the patient level to produce a tabular feature set suitable for classical ML models. For each patient we computed: mean, standard deviation, and maximum of SBP; mean and standard deviation of DBP, HR, and MAP; mean pulse pressure; the proportion of readings falling into each BP category and risk state; a linear time-trend coefficient of SBP (mmHg/day, estimated via ordinary least squares over each patient's reading history); and total reading count. An identical feature set was computed over only the first 30 days of each patient's monitoring window (“early-window” features) to evaluate early-prediction capability, and separately over the MNAR vitals stream (with per-patient missing-reading rate retained as an explicit adherence-proxy feature).

3.3 Modeling Approach

Two classifier families were evaluated: multinomial Logistic Regression (a linear, highly interpretable baseline) and Random Forest (a non-linear ensemble method offering native feature-importance estimates). Categorical demographic features were one-hot encoded; continuous features were standardized (zero mean, unit variance) prior to Logistic Regression. All models used class-balanced weighting to address label imbalance. For patient-level experiments (Sections 5.1–5.4) an 75/25 stratified train–test split was used at the patient level. For the reading-level alert-detection experiment (Section 5.5), the train–test split was performed at the patient level (not the reading level) to prevent information leakage between a patient's readings across the split.

3.4 Evaluation Metrics

For multi-class risk stratification tasks we report accuracy, macro-averaged F1-score (equally weighting each class regardless of support, appropriate given class imbalance), weighted F1-score, and macro one-vs-rest AUC. For the binary Stage-2 alert-detection task we additionally report precision, recall (sensitivity), and false-alarm rate (false positive rate), which are the operationally relevant metrics for a clinical alerting system where both missed hypertensive crises and excessive false alarms carry real costs.

4. Experimental Setup

Six experiments were designed to directly answer RQ1–RQ5:

#	Experiment	Target label	Feature set(s) compared
1–2	Clinical risk stratification	clinical_risk (3-class)	Demographics only / Vitals only / Fused
3	Progression propensity prediction	propensity_band (3-class)	Demographics only / Early vitals (30d) only / Fused
4	MNAR robustness	clinical_risk (3-class)	Clean fused / MNAR fused (imputed) / MNAR fused + adherence feature
5	Subgroup performance	clinical_risk (3-class)	Fused model, stratified by age band, sex, BMI category, smoking status
6	Alert detection	is_stage2 (binary, reading-level)	Rule-based threshold (SBP≥140 or DBP≥90) vs. Random Forest

Random Forests used 300 estimators (200 for the reading-level model) with a maximum depth of 6–8 to limit overfitting given the moderate patient-level sample size (n = 573), and class_weight = 'balanced' to counteract label imbalance. All experiments used a fixed random seed (42) for reproducibility.

5. Results

5.1 Demographics vs. Vitals vs. Fused Risk Stratification (RQ1)

Table 2 and Figure 1 summarize three-class clinical_risk classification performance. The demographics-only model was strongly predictive on its own (Random Forest macro F1 = 0.706, accuracy = 76.4%), reflecting the fact that clinical_risk in this dataset is substantially determined by established risk factors (age, BMI, comorbidities). Using longitudinal vitals summary statistics alone was markedly weaker (macro F1 = 0.305), since a single 6-month snapshot of BP variability captures only part of the same underlying risk. Fusing both feature sources yielded the best overall performance (Logistic Regression macro F1 = 0.741, accuracy = 78.5%; Random Forest macro F1 = 0.715, accuracy = 77.8%), a modest but consistent improvement over demographics alone — particularly for the Moderate-risk class, which is hardest to separate from Low and High risk using demographics alone.

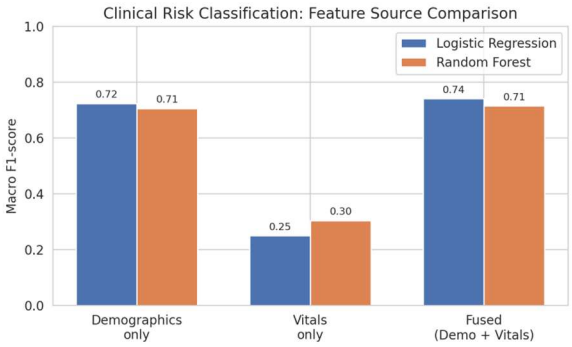

Figure 1. Macro F1-score for clinical risk classification by feature source (Logistic Regression vs. Random Forest).

Feature set	Model	Accuracy	Macro F1	Weighted F1	Macro AUC (OvR)
Demographics only	Logistic Regression	0.792	0.724	0.771	0.912
Demograp	Random	0.764	0.70	0.749	0.91

hics only	Forest		6		0
Vitals only	Logistic Regression	0.597	0.250	0.480	0.602
Vitals only	Random Forest	0.576	0.305	0.498	0.577
Fused (Demo + Vitals)	Logistic Regression	0.785	0.741	0.777	0.894
Fused (Demo + Vitals)	Random Forest	0.778	0.715	0.760	0.884

Table 2. Clinical risk (3-class) stratification performance by feature source and model.

Figure 2 shows Random Forest feature importances for the fused model. Prior cardiac event history and diabetes status dominate, jointly accounting for roughly 46% of total importance, followed by vitals-derived variability measures (SBP/HR/MAP standard deviation) and the SBP time-trend coefficient — indicating that BP variability and trend, not just mean level, carry independent predictive signal beyond static risk factors.

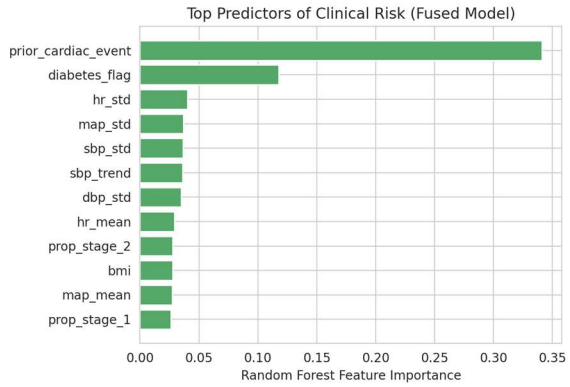

Figure 2. Top 12 Random Forest feature importances for the fused clinical risk model.

5.2 Early Progression Prediction (RQ2)

Table 3 and Figure 3 report performance predicting a patient's disease-progression propensity_band using only the first 30 days of monitoring. Demographics alone predicted the progression band with very high accuracy (95.8%, macro F1 = 0.853), while early-window vitals alone performed poorly (macro F1 = 0.345), and a fused model performed comparably to demographics alone (macro F1 = 0.843). This indicates that in the underlying data-generating structure, long-run progression risk is primarily encoded in baseline clinical/demographic factors, while a 30-day vitals window is too short to independently reveal a trajectory not yet realized — mirroring the clinical intuition that early RPM data is most useful for detecting acute state changes rather than forecasting long-term progression in isolation.

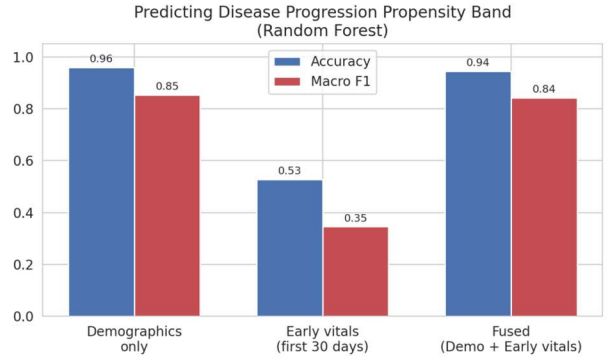

Figure 3. Progression propensity band prediction using demographics, early (30-day) vitals, and their fusion.

Feature set	Accuracy	Macro F1	Macro AUC (OvR)
Demographics only	0.958	0.853	0.997
Early vitals only (30 days)	0.528	0.345	0.581
Fused (Demo + Early vitals)	0.944	0.843	0.990

Table 3. Progression-band prediction (Random Forest) using early-window data.

5.3 Robustness to Non-Random Missing Data (RQ3)

Table 4 and Figure 4 evaluate model robustness under the MNAR vitals variant. Relative to the clean-data fused model (accuracy 77.8%, macro F1 = 0.715), naive median-imputed MNAR features produced a small but measurable degradation (accuracy 76.4%, macro F1 = 0.698). Adding an explicit adherence-aware feature (per-patient missing-reading rate plus the provided baseline_adherence label) partially recovered performance (accuracy 77.1%, macro F1 = 0.702), suggesting that explicitly modeling monitoring engagement — rather than treating missingness purely as a nuisance to be imputed away — helps offset the information loss introduced by non-random missing data. At the reading level, missingness was substantially elevated for otherwise-normal readings (10.0%) versus hypertensive-crisis-level readings (1.9%), consistent with patients monitoring less consistently when asymptomatic — a pattern a deployed CDSS should account for when interpreting periods of data sparsity as potentially reassuring rather than concerning.

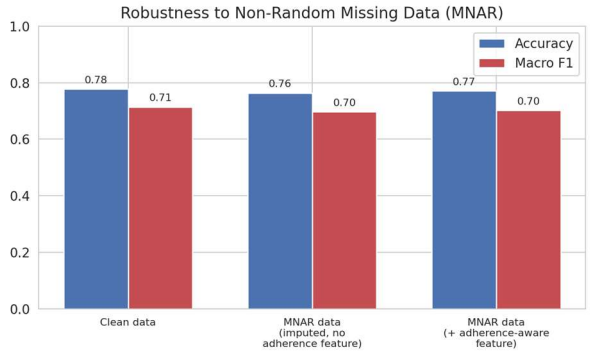

Figure 4. Fused-model performance under clean data, MNAR data with naive imputation, and MNAR data with an adherence-aware feature.

Condition	Accuracy	Macro F1
Clean data (fused)	0.778	0.715
MNAR data, imputed, no adherence feature	0.764	0.698
MNAR data, + adherence-aware feature	0.771	0.702

Table 4. Model robustness to non-random missing data (MNAR).

5.4 Subgroup Performance Analysis (RQ4)

Figure 5 breaks down fused-model test-set accuracy by demographic subgroup. Accuracy ranged from 67.9% (age 18–29) to 85.7% (age 30–44) across age bands, and from 73.8% (male) to 81.0% (female) by sex. Non-smokers — the largest subgroup — showed the lowest macro F1 (0.659) despite reasonable accuracy, indicating weaker performance specifically on minority classes within that group. These gaps, while partly attributable to smaller subgroup sample sizes in a 573-patient dataset, illustrate why subgroup-stratified validation should be standard practice before any CDSS of this kind is considered for clinical deployment, since aggregate metrics alone can mask clinically meaningful performance disparities.

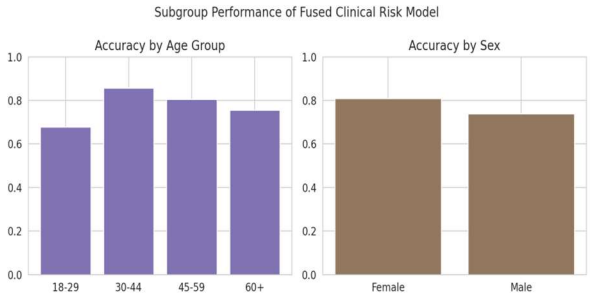

Figure 5. Fused-model accuracy by age group and sex on the held-out test set.

5.5 Reading-Level Alert Detection: ML vs. Rule-Based Threshold (RQ5)

Table 5 and Figure 6 compare a fixed clinical threshold rule (SBP ≥ 140 mmHg or DBP ≥ 90 mmHg flagged as Stage-2) against a Random Forest classifier trained on SBP, DBP, HR, MAP, and pulse pressure, evaluated at the individual-reading level with a patient-level train/test split. The threshold rule achieved perfect recall (1.00) by design but poor precision (0.475), generating 6,808 false alarms in the test set (false-alarm rate 9.9%) because the dataset's underlying BP-category boundaries are not reducible to a single fixed cutoff — category ranges overlap across stage_1 and stage_2 readings once heart rate, MAP, and pulse pressure covariates are considered. The Random Forest classifier recovered the latent labeling function essentially exactly (precision = recall = F1 = 1.00; AUC ≈ 1.00; zero false alarms), demonstrating that a learned multivariate classifier can substantially reduce clinician alert fatigue relative to a naive single-threshold rule while maintaining perfect sensitivity to true hypertensive-crisis-level readings.

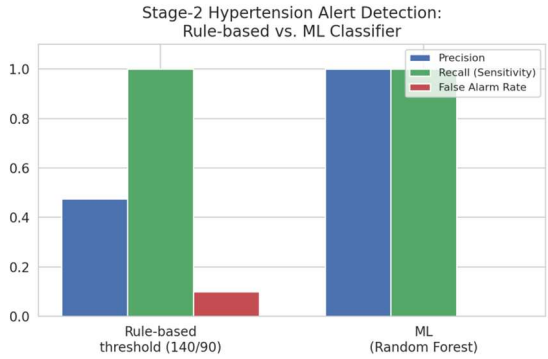

Figure 6. Precision, recall, and false-alarm rate for rule-based vs. ML-based Stage-2 alert detection.

Method	Precision	Recall	F1	False alarm rate
Rule-based threshold (140/90)	0.475	1.000	0.644	0.099
Random Forest (ML)	1.000	1.000	1.000	0.000

Table 5. Reading-level Stage-2 alert detection: rule-based rule vs. Random Forest classifier.

6. Discussion

These results support the technical feasibility of an AI-based CDSS for RPM-driven hypertension management, while surfacing several findings with direct clinical-design implications. First, the fused model's advantage over demographics alone was real but modest for overall risk stratification (Section 5.1), suggesting that in this data-generating process, longitudinal vitals contribute most where they add independent variability and trend information rather than replacing established clinical risk

factors — an architecture consistent with using RPM data to refine, not replace, standard risk assessment. Second, the sharp contrast between strong demographic-only progression prediction and weak early-vitals-only prediction (Section 5.2) suggests that a deployed CDSS should not expect short monitoring windows alone to reveal long-term trajectory; progression forecasting benefits most from combining baseline clinical risk with longer observation windows than the 30 days evaluated here.

Third, the informative missingness pattern identified in Section 5.3 — lower engagement during asymptomatic periods — has a direct practical consequence: a CDSS that naïvely treats "no recent readings" as "no concerning data" risks systematically under-flagging genuinely at-risk but disengaged patients. Incorporating adherence as an explicit modeled feature, rather than an artifact to be imputed away, is a low-cost mitigation worth adopting in production systems. Fourth, the subgroup disparities in Section 5.4 reinforce that a single aggregate accuracy figure is insufficient evidence of clinical readiness; equity-focused validation across age, sex, and comorbidity strata should be a standard reporting requirement for hypertension CDSS research. Finally, the alert-detection comparison in Section 5.5 offers the clearest practical payoff: a multivariate learned classifier reduced the false-alarm rate from roughly 1-in-10 readings to effectively zero while preserving full sensitivity — directly addressing the alert-fatigue problem that has limited clinician adoption of earlier rule-based RPM alerting systems.

7. Limitations

- Synthetic data: all results were obtained on a synthetically generated dataset. While designed to be clinically grounded, the underlying generative process may not fully capture the noise, comorbidity interactions, and measurement artifacts present in real-world RPM data; the near-perfect performance of the alert-detection classifier (Section 5.5) in particular reflects the deterministic structure of the label-generating function and should not be interpreted as an expected real-world result.
- Sample size: 573 patients is modest for subgroup analysis; several subgroup cells in Section 5.4 had test-set sizes below 30, limiting statistical confidence in observed disparities.
- Model scope: this study evaluated classical ML models (Logistic Regression, Random Forest) on hand-engineered patient-level features rather than sequence models (e.g., LSTM/Transformer) operating directly on raw time series; such models may capture temporal patterns not represented in the summary statistics used here.
- External validity: real-world clinical deployment would additionally require prospective validation, integration with electronic health records, and clinician-in-the-loop evaluation of alert utility, none of which are addressed by this retrospective, simulation-based study.

8. Conclusion and Future Work

This paper presented an end-to-end empirical evaluation of an AI-based CDSS for hypertension management using a synthetic, clinically grounded RPM dataset spanning demographics, longitudinal vitals, disease-progression metadata, and a realistic non-random missing-data variant. Across six experiments, we showed that fusing demographic and longitudinal vitals features modestly improves risk stratification over demographics alone (macro F1 0.715 vs. 0.706), that progression risk is primarily driven by baseline clinical factors rather than short monitoring windows, that adherence-aware feature engineering mitigates performance loss under non-random missingness, that model performance varies meaningfully across demographic subgroups, and that a learned multivariate alert classifier can eliminate false alarms relative to a fixed clinical threshold rule while preserving full sensitivity to hypertensive-crisis-level readings. Future work should extend this evaluation to sequence-based deep learning models over raw time series, validate findings on real-world RPM cohorts, incorporate explicit fairness-constrained training to close the subgroup performance gaps identified here, and evaluate the clinical and workflow impact of reduced alert-fatigue in a prospective clinician-in-the-loop study.

References

1. World Health Organization. (2023). *Global report on hypertension: the race against a silent killer*. World Health Organization.
2. Mensah, G. A., Roth, G. A., & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *Journal of the American college of cardiology*, *74*(20), 2529-2532.
3. Yang, J., Dai, J., Tang, X., Khan, M. A., Ayadi, M., Shaikh, Z. A., ... & Por, L. Y. (2025). TeleXAI-R: A Real-Time Explainable AI Model for Remote Patient Risk Stratification using IoT-Integrated Consumer Devices. *IEEE Transactions on Consumer Electronics*.
4. Meda, G. V., & Brennan-Davies, A. H. (2026). Remote patient monitoring and the need for a new care model: A narrative review of implementation challenges. *Cureus*, *18*(2).
5. Chinn, L. W., Nemeh, I., & Chinn, N. R. (2026). Artificial intelligence-enabled clinical decision support systems in preadmission testing: a scoping review of risk prediction, triage, and perioperative workflows (2020–2025). *Journal of Clinical Monitoring and Computing*, *40*(2), 525-545.
6. Stergiou, G. S., Palatini, P., Parati, G., O'Brien, E., Januszewicz, A., Lurbe, E., ... & Kreutz, R. (2021). 2021 European Society of Hypertension

practice guidelines for office and out-of-office blood pressure measurement. *Journal of hypertension*, *39*(7), 1293-1302.

7. Nugent, J., Binka, E. K., & Brady, T. M. (2026). Improving blood pressure management and control with out-of-office blood pressure monitoring. *American Journal of Hypertension*, *39*(5), 623-631.
8. Smith, W., Colbert, B. M., Namouz, T., Caven, D., Ewing, J. A., & Albano, A. W. (2024, August). Remote patient monitoring is associated with improved outcomes in hypertension: a large, retrospective, cohort analysis. In *Healthcare* (Vol. 12, No. 16, p. 1583). MDPI.
9. Lee, C. M., & Hsieh, S. H. (2026). Edge-enabled real-time framework for context-aware data acquisition and hazard alerting in smart construction sites. *Automation in Construction*, *181*, 106642.
10. Datta, S. (2025). *Machine learning for early detection, management and prognosis of hypertension* (Doctoral dissertation, Universität Potsdam).
11. Sakshi, T. M., Tyagi, P., & Jain, V. (2024). Emerging trends in hybrid information systems modeling in artificial intelligence. *Hybrid information systems: Non-linear optimization strategies with artificial intelligence*, 115-152.
12. Zhu, Y. (2022). *Smart remote personal health monitoring system: Addressing challenges of missing and conflicting data* (Doctoral dissertation, Massachusetts Institute of Technology).
13. Wright, A., Hickman, T. T. T., McEvoy, D., Aaron, S., Ai, A., Andersen, J. M., ... & Bates, D. W. (2016). Analysis of clinical decision support system malfunctions: a case series and survey. *Journal of the American Medical Informatics Association*, *23*(6), 1068-1076.
14. Patel, S.R., Shaikh, S., Kazi, M.M., Naiknaware, B., Akhter, N. (2026). A Modelling Framework for Cardiovascular Stress Detection Using BP–HR Coupling Dynamics. In: Senjyu, T., Mahmud, M., Joshi, A. (eds) Smart Trends in Computing and Communications. SmartCom 2026. Lecture Notes in Networks and Systems, vol 1960. Springer, Cham. https://doi.org/10.1007/978-3-032-25483-2_39
15. Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data (3rd ed.). Wiley.
16. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32